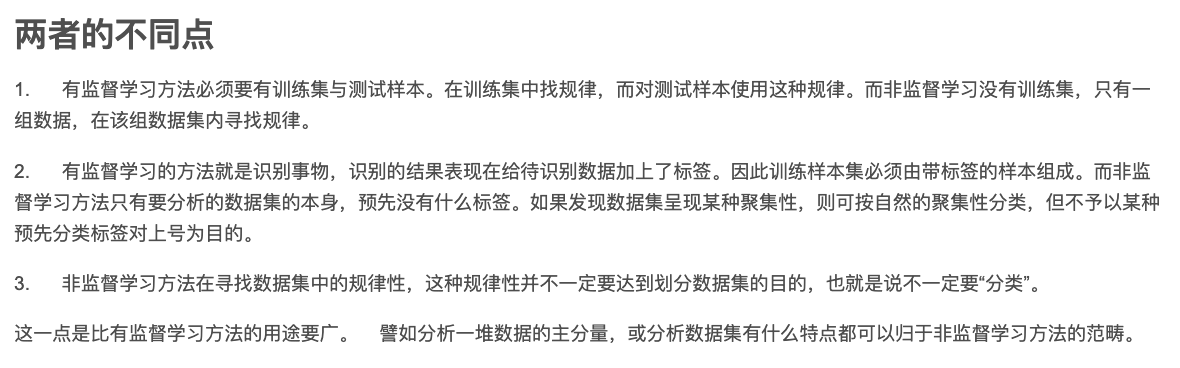
第一章

监督学习：训练集包括特征和标记信息。监督学习就是最常见的分类问题，通过已有的训练数据去训练去训练得到一个模型，再利用这个模型将所有的输入映射为相应的输出。比如决策树，神经网络都是监督学习。

无监督学习：训练集没有标记信息。



分类与回归的原理与区别

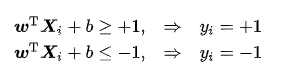
**回归与分类的根本区别在于输出空间是否为一个度量空间**

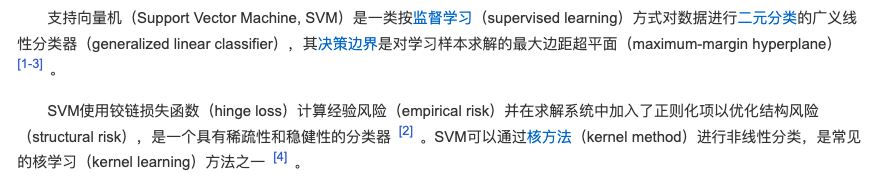
回归，指研究一组随机变量(Y1 ，Y2 ，…，Yi)和另一组(X1，X2，…，Xk)变量之间关系的统计分析方法



第六章

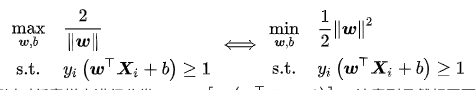
支持向量：距离超平面最近的几个训练样本点并且使下列等式成立





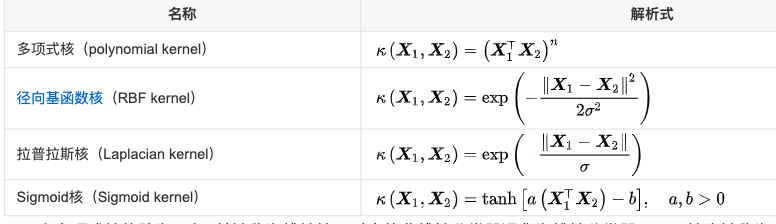
间隔：两个异类支持向量到超平面的距离之和

**目标**是找到具有“最大间隔”的划分超平面（凸二次规划问题）

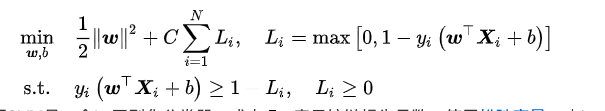


核函数：[支持向量机](https://baike.baidu.com/item/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA/9683835" \t "/Users/mac/Documents\\x/_blank)通过某非线性变换 φ( x) ，将输入空间映射到高维特征空间。支持向量机的求解用到内积运算φ( xi)Tφ( xj) ，而在低维输入空间又存在某个函数 K(x, x′) ，它恰好等于在高维空间中这个内积，即K( x, x′) =<φ( x) ⋅φ( x′) > 。这样的函数 K(x, x′) 称为核函数。

常用核函数：



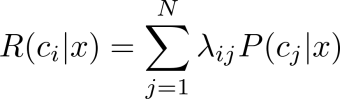
软间隔：**允许某些样本不满足约束**，因此在优化目标中引入损失函数。常见的有hinge损失，指数损失，对率损失。下式Li为“**松弛变量**”。



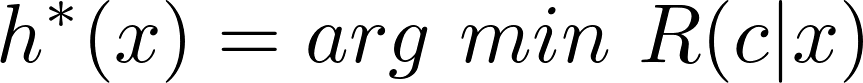
**正则化**：可理解为一种“罚函数法”，即对不希望得到的结果施以惩罚，从而使得优化过程逐渐趋向于希望目标，从贝叶斯估计的角度来看，正则化项可认为是提供了模型的先验概率。

第七章

**基本原理**：基于样本先验概率，利用贝叶斯公式计算后验概率，选择具有最大后验概率的类作为该样本的分类。

条件风险(将样本错误分类为ci所产生的期望损失）：

目标是寻找一个判断准则h：X->Y，能最小化条件风险。这就产生了**贝叶斯判定准则**：**为最小化总体风险，只需在每个样本上选择那个能使条件风险R（c|x）最小的类别标记**，即



h\*称为**贝叶斯最优分类器**。

**估计类条件概率的常用策略**：先假定类条件概率具有某种确定的概率分布形式，再基于训练样本对概率分布的参数进行估计。

**极大似然估计**：是试图在参数值所有可能的取值中，找到一个使目标数据出现的“可能性”最大的值。

**朴素贝叶斯分类器**：属性条件独立性假设。

EM算法：解决训练样本不完整的问题，即隐变量估计问题，非梯度优化方法。

第一步是期望（E）步，利用当前估计的参数值来计算对数似然的期望值；第二步是最大化（M）步，寻找能使E步产生的似然期望最大化的参数值。然后，新得到的参数值重新被用于E步，直至收敛到局部最优解。

第八章

集成学习：通过构建并结合多个学习器来完成学习任务

同质：只包含同种类型的个体学习器

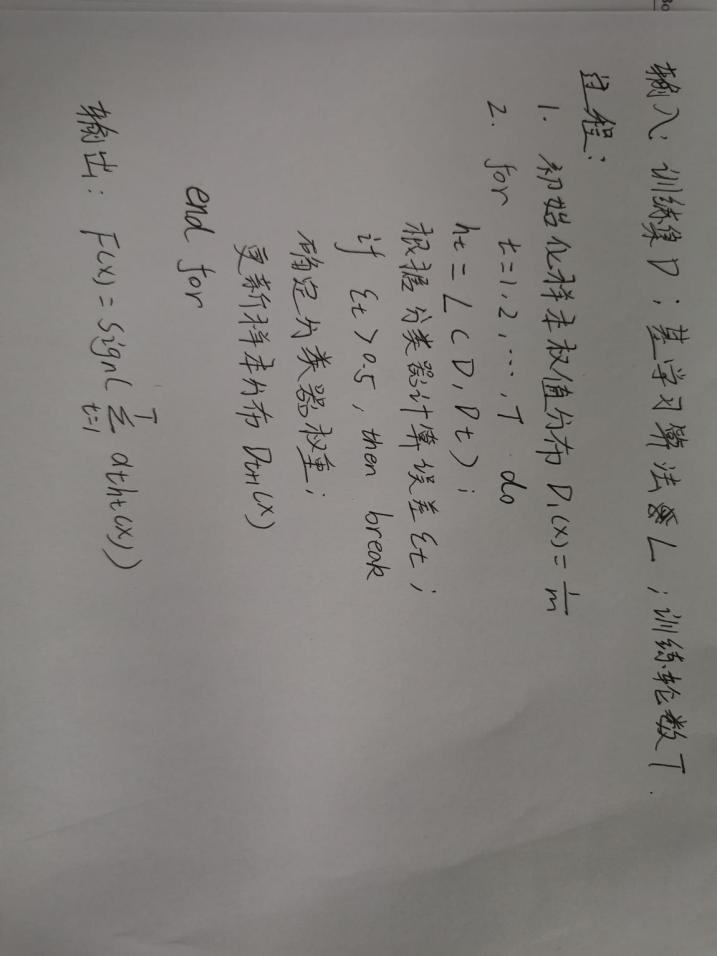
异质：包含不同类型的个体学习器

集成个体应“好而不同”，即“准确性”和“多样性”。

Boosting

个体学习器间存在强依赖关系，必须串行生成的序列化方法。主要关注降低偏差。代表算法为AdaBoost（只适合二分类）。

基本原理：先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，然后基于调整后的样本分布来训练下一个基学习器，如此重复进行，直至基学习器数目达到事先指定的值T，最终将这T个基学习器进行加权组合。



基学习器对特定数据分布进行学习的方法：

1. 重赋权法：在每一轮训练中，根据样本分布为每个训练样本重新赋予一个权重。
2. 重采样法：在每一轮训练中，根据样本分布对训练集重新进行采样，再用重采样而得的样本集对基学习器进行训练。**可以获得“重启动”机会以避免训练过程过早停止。**

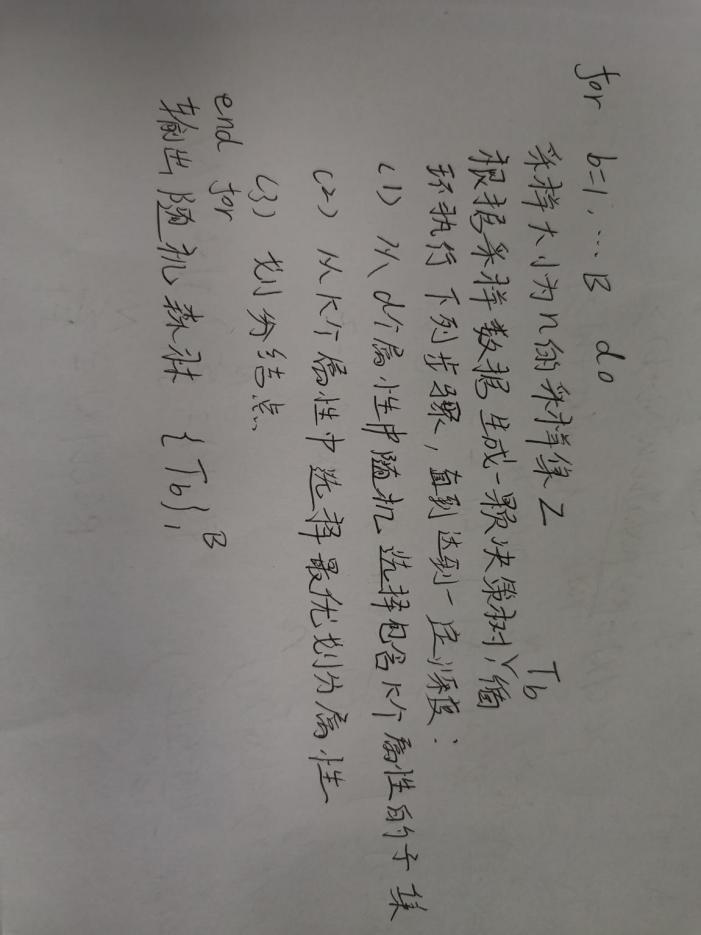
**Bagging**

个体学习器之间不存在强依赖关系，可同时生成的并行化方法。主要关注降低方差。基于**自助采样法**。

基本原理：采样出T个含m个训练样本的采样集，然后基于每个采样集训练出一个基学习器，再将这些基学习器进行结合（常采用简单投票法）

随机森林（RF）

以决策树为基学习器构建Bagging集成的基础上，进一步在决策树的训练过程中引入随机属性选择。



优点：

1. 训练可以并行化
2. 最优属性划分选择时是从属性子集中进行选择，训练开销减小
3. 基学习器的多样性不仅来自样本扰动，还来自属性扰动，泛化性能进一步提升

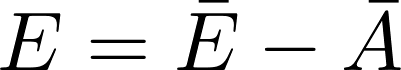
缺点：

1. 在噪音比较大的样本集上面，容易过拟合。
2. 取值划分比较多的特征容易对RF决策产生更大的影响从而影响拟合结果。

结合策略：

1. 平均法：包括简单平均和加权平均
2. 投票法：包括绝对多数投票法，相对多数投票法，加权投票法
3. 学习法。针对训练数据很多，典型代表为Stacking

多样性

误差-分歧分解：，个体学习器准确性越高、多样性越大，则集成越好。

第九章

聚类定义：将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。

性能度量：

1. 外部指标：将聚类结果与某个“参考模型”进行比较
2. 内部指标：直接考察聚类结果

距离计算：闵可夫斯基距离，无序属性采用VDM。

A.原型聚类：

1. k均值算法
2. 学习向量量化（LVQ）：假设样本带有类别标记
3. 高斯混合聚类：采用概率模型来表达聚类原型

B.密度聚类

考察样本之间的可连接性，代表算法DBSCAN

C.层次聚类

代表算法AGNES，采用自底向上聚合策略。

K均值算法

最小化平方误差，刻画了簇类样本围绕簇均值向量的紧密程度。